Amir Zeldes, Anke Lüdeling and Hagen Hirschmann
Humboldt-Universität zu Berlin
amir.zeldes@rz.hu-berlin.de, anke.luedeling@rz.hu-berlin.de, hagen_h@yahoo.com

What's Hard? Quantitative Evidence for Difficult Constructions in German Learner Data

1. Introduction

Our study is concerned with the identification of 'difficult' structures in the acquisition of a foreign language, which will shed light on theoretical considerations of L2 processing. We argue that – compared to simple vocabulary items or abstract syntactic patterns – structures that contain lexical material as well as categorial variables are especially difficult to acquire. The difficulty level for particular patterns is shown to depend on surface invariability but not on the syntactic categories within which target patterns are embedded. As an example we study the distribution of certain structures which are underused by L2 German learners.

The question "what is difficult for a language learner?" can be addressed using several kinds of data, including learner corpora (e.g. error analysis and over/underuse data, for an overview see Granger et al. 2002), elicitation data, or psycholinguistic studies. Here we focus on corpus data. Previous corpus studies focusing on learner difficulties have examined token and type frequencies in order to calculate vocabulary richness measures, such as lexical density as an index of learner competence (Halliday 1989, Laufer & Nation 1999, and many others). However, lexical frequencies do not tell us what constructions are difficult for learners beyond individual lexemes, nor why. Many other studies (examples are Borin & Prütz 2004 or Westergren-Axelsson & Hahn 2001) focus on interference errors due to the learners' native language (or other learned languages) by comparing learners with a certain L1 to native speakers. Yet in order to establish explanations for difficulties in L2 acquisition independent of a learner's native tongue, we must examine the distributions in native and learner data of e.g. lexemes, collocations, colligations (cf. Stefanowitsch & Gries 2003) and syntactic structures, across learners' linguistic backgrounds. We take the stance that L1-independent underuse phenomena are due to learners either not acquiring patterns, or else avoiding their use despite familiarity with them, in both cases indicating increased difficulty.

2. Data

The data for this study comes from the Falko corpus (**F**ehler**a**nnotiertes **L**erner**ko**rpus des Deutschen als Fremdsprache), which consists of texts from advanced learners of German and control data from German L1-speakers (Lüdeling et al. 2008), allowing contrastive interlanguage analyses. The corpus is stored in a multi-layer model searchable at various levels of annotation. In order to diminish the possibility that the learners are simply unfamiliar with the items in question, we examine only advanced learners and focus on frequent, prevalent patterns. To filter out interference from the learners' L1 and other foreign languages we examine data from speakers of five different L1s: Danish (da), English (en), French (fr), Polish (pl) and Russian (ru), with diverse language education. Using this data, we examine the normalized frequencies of all word form types and part-of-speech *n*-grams in order to find the most significant cases of underuse. Here we focus on two particularly striking cases found in this way, involving reflexives and adverb chains. Use of the reflexive pronoun *sich* can be difficult for learners (Mode 1996), since they must learn not only which verbs and senses require it, but also correctly position it either after the verb in a main clause (1), after a complementizer (2) or subject (3) in subordinate clauses, or initially in an

infinitive phrase (4). Treating the usage of *sich* as a random variable and using a test of equal proportions our data shows very significant underuse of *sich* in both learners in total vs. natives, and each learner dataset grouped by native language vs. natives.

1. *sie entscheiden **sich** meistens für die Firma*
   they decide [refl] usually for the firm
   they usually decide for the firm
2. *dass **sich** die Frauen überfördert fühlen*
   that [refl] the women over-challenged feel
   that the women feel they can't cope
3. *Als die Stadt **sich** ändert*
   as the city [refl] changes
   As the city changes
4. ***sich** ihren Mann auszusuchen*
   [refl] her husband choose
   to choose her (own) husband

L1	natives	learners	da	en	fr	pl	ru
f(*sich*)	.011697	.005910	.006283	.006291	.006930	.007170	.005435
tokens	74280	88736	15593	21600	7786	18100	11203
p-val.		< 2.2e-16	< 3.314e-9	< 8.518e-12	< 1.849e-4	< 1.595e-7	< 3.465e-9

Learners use *sich* about half as often as natives, independent of their L1, even though *sich* is the 17th most common word form in the corpus overall, so it can be assumed that the learners are familiar with it. The examined L1s are quite diverse with regard to the morphosyntax of reflexives (e.g. enclitic or not, position relative to the verb, variability depending on the finite verb's person), yet four of them have similar reflexives (da. *sig*, fr. *se*, pl. *się*, ru. *-sja*). This reduces the likelihood of interference accounting for the underuse phenomenon. Additionally, since interference by definition depends on the learner's native language, we would expect some statistical differences in the underuse patterns between learners with different L1s if interference were a factor (more or less underuse depending on the amount or type of interference). However, the frequency of *sich* in all five learner datasets does not differ significantly (p-val. of .4478 in a 5-way test of equal proportions). Another possible difficulty could be word order complexity of *sich* in relative/infinitive clauses (1-4 above). Yet the data shows *sich* is similarly underused in all syntactic environments, with learner/native normalized frequency ratios of .54 for main clauses, .55 for subordinate clauses and .62 for infinitive clauses, with no significant difference (p-val. of .354 in Pearson's chi-squared test). We therefore conclude that *sich* is similarly underused by our learners independently of their L1 and the embedding clause type.

By contrast, learners do use *sich* more often in certain less variable contexts, such as when the subject is the generic pronoun *man* 'one' (5), despite the fact that *man* itself is not in overuse (an insignificant underuse ratio of ~.95). In these cases the word order in (2) is ungrammatical and only (3) is possible, i.e. *man sich*. This recurring surface pattern is not underrepresented in the learner data (an insignificant underuse ratio of ~.9, cf. row 1 of the table below). Similarly, combinations of *sich* with *lassen* 'allow, let' (6) are also frequent despite an underuse ratio of ~.56 for *lassen*, actually being overused in datasets from three learner L1s and overall (overuse ratio above 1.5 in row 2, though not statistically significant):

5. *Wenn **man sich** bemüht*
   if one [refl] exert
   If one makes the effort

6. *Anhand dieses Beispiels* ***läßt sich*** *erschließen*
   using this example allows [refl] conclude
   Using this example it is possible to conclude

pattern \ L1	learners/ natives	natives	learners	da	en	fr	pl	ru
man* + *sich	.9079	.000563	.000512	.000834	.000509	.001027	.000276	.000089
lassen* + *sich	1.5359	.000078	.000121	.000064	.000185		.000110	.000178
ADV + ADV	.452	.01285	.00581	.01051	.00611	.00616	.00309	.00285
ADV x 3	.265	.00182	.00048	.00109	.00051	.00038	.00011	.00026

In addition to the lexical underuse data above, we also compare frequencies of part-of-speech chains (PoS bigrams and trigrams) in the same corpus. The PoS chains most underrepresented in all examined learner datasets contain two or three consecutive adverbs (and some particles tagged as adverbs, due to the STTS tagset used), with p-value < 2.2e-16 for the bigrams and 1.776e-14 for the trigrams. To explain this phenomenon we examine the 30 most frequent pairs of adverbs qualitatively, since the total amount of chains is too small to evaluate statistically. In order to abstract beyond specific lexical adverb bigrams we divide the chains into four main categories: I. the adverbs belong to different phrases (a 'quasi-pair'; (7) and (8)); or else the adverbs belong to the same phrase which is either II. left-headed (9), III. right-headed (10) or IV. lexicalized (11).

7. *Es ist* [***doch***] [***auch***] *statistisch belegt, dass*
   it is indeed also statistically proven that
   Furthermore, it is indeed statistically proven that
8. *die* (...) *haben* [***schon***] [[***ziemlich*** *viele*] *Lebenserfahrungen*]
   they have already quite many life-experiences
   they already have quite a lot of life experience
9. *ein Kampf, dass bis* [***heute noch***] *andauert*
   a fight that until today still endures
   a fight which has lasted until today
10. *wo es* (...) [[***viel mehr***] *Arbeitsplätze*] *gibt*
    where it much more jobs gives
    where there are many more jobs
11. *und* [***immer noch***] *kann man eine unzufriedenheit spüren*
    and always still can one a discontentment sense
    and still one can sense some discontentment

In category I, we notice a difference between the use of pair types whose elements are sentence- or VP-modifying adverbs (forming two adverbial phrases), as in (7), and those whose second element is a modifier to an adjective phrase or a DP (as an adverbial particle), as in (8). Structures like (7) are very rare in the learner data, whereas structures like (8) seem not remarkably underrepresented. We explain these findings by the different variability of the structures themselves: in (7) the second of the two adverbs (*auch*) can be moved to the initial position of the sentence (*Auch ist es doch statistisch belegt, dass*), or additional elements/phrases can be inserted directly before or after it. Its position is therefore relatively flexible. In (8) *ziemlich* is bound to the adjective phrase with *viele* as a head, it cannot be moved in the sentence without its DP, and no element can be inserted between *ziemlich* and *viele*; its position is fixed. We argue that the differences in frequency are due to differences in the variability of the structures – learners seem to either not acquire topologically flexible elements or be insecure as to where to place them and opt to avoid them.

The single phrase categories II-IV show different patterns. The left-headed phrases in category II (e.g. *heute noch* 'still today' in ex. (9)) are (not surprisingly) the least frequent, with too little data to draw any conclusions from. Category III (right-headed phrases like *viel mehr* 'much more' in ex. (10)) is similarly attested for learners and natives, which can be explained by its easy to learn and topologically fixed surface structure. This structure is similar to that of adverbs followed by attributive adjectives (e.g. use of the intensifier *sehr* 'very', in [*eine* [***sehr liebenswerte***] *Gattin*] 'a **very loveable** wife'), which show no statistically significant under- or overuse at all. This may be because their structure is even easier to learn than the one in (8): they have a fixed pattern DP[DET NP[AP[ADV A] N]] with an invariable topological structure.

The lexicalized pairs in category IV (e.g. *immer noch* 'still', ex. (11)) have to be analysed as single units (with no internal structure). Most of these phrases can be at least partly expressed by just one word (*immer noch* → *noch* 'still'), which learners may choose to use instead. We do not find systematic underuse or overuse in all of these cases – such lexical units can apparently be learned like any other, and frequent ones appear to be better represented in many learner groups (e.g. *immer noch* in the Danish, English, and French subcorpora).

3. Discussion

The learner difficulties examined in our study, as identified by underuse statistics, suggest that complex constructions with variable surface forms, such as mobile reflexive pronouns and non-lexicalized adverb chains, hinder effective acquisition of native-like language production. Invariable, frequently recurring patterns, such as lexicalized chains and combinations like reflexive + *man* or *lassen*, facilitate the use of the corresponding constructions. These results conflict with an algebraic model of grammar that might predict that all reflexive verbs and adverb chains are equally likely to be learned, regardless of lexemes (certain adverbs or verbs) or embedded/embedding constructions (*man* as a subject); but they also conflict with models based solely on input frequency. Diverging from the target language distribution, learners seem to filter out reflexives and multiple adverbs in the native usage they are exposed to, but less so when these are embedded in recurrent patterns. This points to a quantitative destructive effect of surface form variability on the learnability of complex structures, possibly connected to processing considerations in the absorption of items in the mental lexicon.

4. References

Borin, L./Prütz, K. (2004) New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In: Aston, G./Bernardini, S./Stewart, D. (eds.) *Corpora and Language Learners*, 67-88. Amsterdam: John Benjamins.

Granger, S./Hung, J./Petch-Tyson, S. (eds.) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Stefanowitsch, A./Gries, S. Th. (2003) Collostructions: Investigating the interaction between words and constructions. *Intl. Journal of Corpus Linguistics* 8(2): 209-243.

Halliday, M.A.K. (1989) *Spoken and Written Language*. Oxford: OUP

Laufer, B./Nation, P. (1999) A vocabulary-size test of controlled productive ability. *Language Testing* 16(1): 33-51.

Lüdeling, A./Dolittle, S./Hirschmann, H./Schmidt, K./Walter, M. (2008) Das Lernerkorpus Falko. *Deutsch als Fremdsprache* 2.

Mode, D. (1996) Zur Stellung des Reflexivpronomens *sich* im deutschen Satz. *Deutsche Sprache* 1/96: 34-53.

Westergren-Axelsson, M./Hahn, A. (2001) The use of the progressive in Swedish and German advanced learner English - a corpus-based study. *ICAME Journal* 25: 5-30.